Jurnal Elektronik Ilmu Komputer Udayana
Volume 10, No 4. May 2022

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Identifikasi Ekspresi Idiomatik Menggunakan *Distributional Semantic Based Approach* dan *Truth Discovery*

Ni Made Yuli Cahyani[a1], AAIN Eka Karyawati[a2], Luh Arida Ayu Rahning Putri[a3], Agus Muliantara[a4], Ida Bagus Gede Dwidasmara[a5], Luh Gede Astuti[a6]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Badung, Bali, Indonesia
[1]yulicahyani1101@gmail.com
[2]eka.karyawati@unud.ac.id
[3]rahningputri@unud.ac.id
[4]muliantara@unud.ac.id
[5]dwidasmara@unud.ac.id
[6]lg.astuti@unud.ac.id

Abstract

Idiomatic expressions are phrases that consist of a sequence of two or more words that have a meaning that cannot be predicted from the meaning of the individual words that compose it. Idiomatic expressions exist in almost all languages but are difficult to extract because there is no algorithm that can precisely decipher the structure of idiomatic expressions, so most rule-based machine translation systems generally translate idiomatic expressions by translating word for word their constituents, but the translation results do not produce the true meaning of the idiomatic expression. Based on this problem, the author tries to do research on the identification of the use of idiomatic expressions in Indonesian sentences. First, the author conducts the sentence classification process using BERT to find out whether the sentence contains idiomatic expressions or not. Furthermore, idiomatic expressions are identified based on distributional semantic based approach and then validated automatically using the Truth Discovery method. From the research conducted, the identification of idiomatic expressions in Indonesian sentences using Distributional Semantic Based Approach and *Truth Discovery obtained an accuracy of 0.82; precision 1.0; recall 0.64 and f1-score 0.78.*

Keywords: *Idiomatic Expressions, BERT, Truth Discovery, Validation, Distribution Semantic*

1. Pendahuluan

Bahasa memegang peranan penting yaitu sebagai alat komunikasi dalam kehidupan sosial masyarakat. Dalam berbahasa, suatu makna tidak hanya dilambangkan dalam satu bentuk bahasa, tetapi juga dapat diungkapkan dalam berbagai bentuk. Bentuk adalah ekspresi makna, sehingga bentuk itu sendiri dapat merangsang penafsiran lebih dari satu makna, salah satu contohnya yaitu dapat dilihat dalam penggunaan idiom. Idiom biasa digunakan dalam kegiatan berkomunikasi sehari-hari yaitu untuk mengungkapkan suatu maksud agar penyampaiannya menjadi lebih menarik atau lebih sopan. Penggunaan idiom itu sendiri sering ditemukan dalam puisi, novel, lirik lagu, surat kabar, majalah atau artikel [1]. Ekspresi idiomatik adalah frasa yang terdiri dari urutan dua kata atau lebih yang memiliki makna yang tidak dapat diprediksi dari makna kata-kata individu penyusunnya. Ekspresi idiomatik ada di hampir semua bahasa dan sulit untuk diekstrak karena tidak ada algoritma yang dapat secara tepat menguraikan struktur ekspresi idiomatik. Identifikasi ekspresi idiomatik adalah masalah yang menantang dengan penerapan yang luas. Mengidentifikasi ekspresi idiomatik sangat penting untuk aplikasi pemrosesan bahasa alami seperti *machine translation, information retrieval* dan sebagainya [2]. Sebagian besar sistem mesin terjemahan berbasis aturan umumnya menerjemahkan ekspresi idiomatik dengan cara menerjemahkan kata demi kata penyusun ekspresi idiomatik, sehingga hasil terjemahan tidak menghasilkan makna yang sebenarnya dari ekspresi idiomatik tersebut.

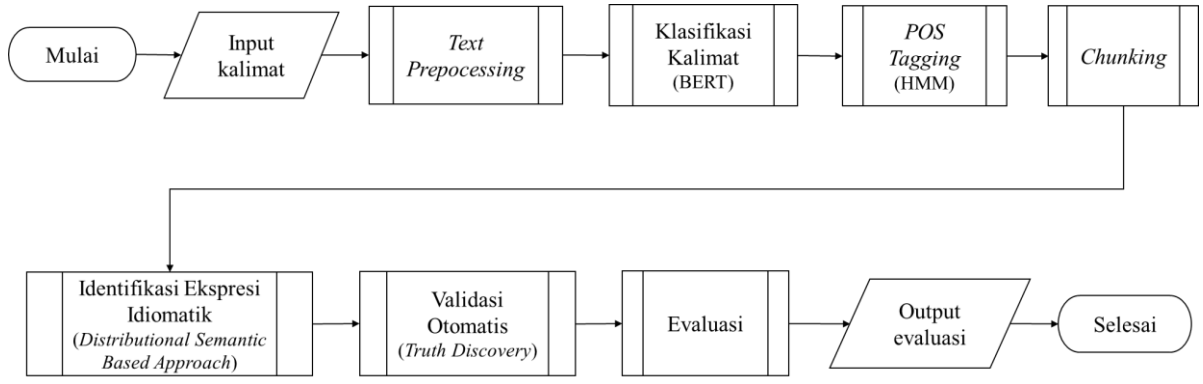
Penelitian tentang identifikasi ekspresi idiomatik bahasa Indonesia belum pernah dilakukan sebelumnya, namun terdapat beberapa penelitian yang serupa dalam bahasa lain. Seperti penelitian [3], pada penelitian ini memperkenalkan pendekatan *semi-supervised* yang menggunakan

representasi terdistribusi dari makna kata untuk menangkap metaforitas. Peneliti menggunakan model *word embedding* untuk mengukur kemiripan semantik antara frasa kandidat dan kumpulan metafora yang telah ditentukan sebelumnya. Penelitian ini memperoleh nilai *precision* 0,5945, *recall* 0,756, *F-score* 0,6657 dan *accuracy* 0,6290. Kemudian terdapat penelitian [4], pada penelitian ini memperkenalkan metode identifikasi metafora pertama yang mengintegrasikan representasi makna yang dipelajari dari data linguistik dan visual dengan menerapkan metode *embedding* kata atau frasa untuk tugas identifikasi metafora. Pada penelitian ini memperoleh nilai *precision* yaitu 0,73, *recall* yaitu 0,80 dan *F-score* 0,76 untuk metode *WordCos* menggunakan *linguistic embeddings*.

Berdasarkan permasalahan di atas, penulis mencoba melakukan penelitian mengenai identifikasi ekspresi idiomatik pada kalimat berbahasa Indonesia. Pertama-tama penulis akan melakukan proses klasifikasi kalimat menggunakan BERT untuk mengetahui apakah kalimat mengandung ekspresi idiomatik atau tidak. Selanjutnya ekspresi idiomatik diidentifikasi berdasarkan pendekatan berbasis semantik distribusi (*Distributional Semantic Based Approach*) yang merupakan kombinasi dari pendekatan pada penelitian sebelumnya yang telah dipaparkan di atas. Metode ini mengidentifikasi ekspresi idiomatik pada tingkat frasa dilakukan dengan menjumlahkan kemiripan semantik antara kandidat dan kumpulan contoh ekspresi idiomatik dengan kemiripan semantik antara kata penyusun frasa kandidat. Kemudian melakukan validasi secara otomatis menggunakan kombinasi algoritma *Sums* dan *Average-Log* yang merupakan algoritma dari metode *Truth Discovery* dengan sumbernya merupakan berbagai macam website yang membahas mengenai ekspresi idiomatik bahasa Indonesia. Diharapkan dengan dilakukannya penelitian ini dapat membantu dalam mengidentifikasi penggunaan ekspresi idiomatik dalam suatu kalimat secara otomatis yang nantinya juga dapat dimanfaatkan untuk membantu mengoptimalkan aplikasi *machine translation* dalam menerjemahkan kalimat yang mengandung ekspresi idiomatik.

2. Metode Penelitian

Penelitian yang dilakukan penulis terdiri dari beberapa tahapan, yaitu data berupa kumpulan kalimat berpola dasar berbahasa Indonesia yang tidak ataupun mengandung ekspresi idiomatik yang telah dikumpulkan dimasukkan ke dalam sistem, setelah itu akan dilakukan *text preprocessing*. Kemudian masuk ke tahap klasifikasi kalimat menggunakan BERT. Selanjutnya kalimat yang diklasifikasikan sebagai ‘kalimat_idiom’ akan masuk ke tahap *POS Tagging* dan *chunking*. Setelah mendapatkan hasil *chunking* berupa frasa, selanjutnya akan masuk ke tahap identifikasi ekspresi idiomatik menggunakan *Distributional Semantic Based Approach*. Setelah itu frasa yang telah diidentifikasi sebagai ekspresi idiomatik akan divalidasi secara otomatis menggunakan metode *Truth Discovery* yaitu kombinasi algoritma *Sums* dan *Average-Log.* Hasil identifikasi yang sudah divalidasi secara otomatis tersebut akan menjadi output akhir dari penelitian ini yang selanjutnya akan masuk ke tahap evaluasi. Secara umum, alur penelitian dapat dilihat pada gambar berikut.



Gambar 1. Alur Umum Penelitian

2.1 Ekspresi Idiomatik Bahasa Indonesia

Ekspresi idiomatik merupakan ungkapan yang maknanya tidak sesuai dengan prinsip komposisionalitas, dan tidak terkait dengan makna bagian [5]. Makna idiomatik adalah makna sebuah satuan bahasa yang menyimpang dari makna leksikal atau makna unsur-unsur pembentuknya. Hal ini berarti suatu idiom tidak dapat diterjemahkan kata per kata tetapi harus dilihat secara utuh dari unsur-unsur pembentuknya. Ekspresi idiomatik yang akan diidentifikasi dalam penelitian ini yaitu idiom yang

p-ISSN: 2301-5373
e-ISSN: 2654-5101

memiliki kategori frasa nomina, frasa verba dan frasa adjektiva yang disusun oleh dua kata. Contoh dari ekspresi idiomatik tersebut adalah sebagai berikut:

Tabel 1. Contoh Ekspresi Idiomatik Bahasa Indonesia

Kategori		Contoh
Frasa Nomina	kata benda + kata benda	Kutu buku
	kata benda + kata sifat	Kuda hitam
	kata benda + kata kerja	Bunga tidur
	kata bilangan + kata benda	Empat mata
Frasa Verba	kata kerja + kata benda	Adu mulut
	kata kerja + kata sifat	Naik pitam
	kata kerja + kata kerja	Jatuh bangun
	kata kerja + kata bilangan	Bermuka dua
Frasa Adjektiva	kata sifat + kata benda	Ringan tangan
	kata sifat + kata sifat	Panjang lebar

2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yaitu data yang sudah tersedia sebelum peneliti memulai penelitian [6]. Terdapat tiga data yang digunakan, dimana data ini bersumber dari media internet, buku ataupun publikasi. Data pertama berupa *Indonesian Manually Tagged Corpus*, yaitu kumpulan kalimat bahasa Indonesia yang telah diberikan tag secara manual. Data ini digunakan pada tahap POS *Tagging*. Contoh data dapat dilihat pada Tabel 2.

Tabel 2. *Indonesian Manually Tagged Corpus*

Contoh Kalimat dan Kelas Katanya
Binatang/NN ini/PR tidak/NEG bisa/MD dibunuh/VB karena/SC masyarakat/NN India/NNP menganggap/VB mereka/PRP suci/JJ ./Z
Jumlah/NN dan/CC harga/NN barang/NN itu/PR masih/MD dirundingkan/VB . /Z
Buaya-buaya/NN itu/PR berukuran/VB panjang/NN antara/IN 40/CD -/Z 50/CD centimeter/NND ./Z

Data kedua berupa kumpulan kalimat berpola dasar berbahasa Indonesia yang mengandung sebuah ekspresi idiomatik dan kalimat berpola dasar berbahasa Indonesia yang tidak mengandung ekspresi idiomatik. Data ini berjumlah 2000 kalimat yang telah dilabeli sebagai kalimat biasa dan kalimat idiom secara manual oleh pakar dan berdasarkan pada buku kamus idiom bahasa Indonesia, dengan jumlah kalimat pada setiap label yaitu 1000 kalimat. Data ini digunakan pada tahap klasifikasi kalimat dan tahap identifikasi ekspresi idiomatik. Contoh data dapat dilihat pada Tabel 3.

Tabel 3. Contoh Kalimat Bahasa Indonesia

indeks	kalimat	kategori	frasa_idiom	validasi
1	Orang tua itu rela membanting tulang demi menyekolahkan ketiga anaknya.	kalimat_idiom	membanting tulang	idiom
2	Ayah adalah tangan kanan Pak Camat.	kalimat_idiom	tangan kanan	idiom
3	Huda suka menjadikan Andik sebagai kambing hitam.	kalimat_idiom	kambing hitam	idiom
4	Gadis itu sedang membaca sebuah buku novel.	kalimat_biasa	none	bukan_idiom
5	Tangan kanan Roni terkena minyak panas saat menggoreng ikan.	kalimat_biasa	none	bukan_idiom
6	Ayah membeli kambing hitam.	kalimat_biasa	none	bukan_idiom

Data berikutnya yaitu data sumber web berupa situs-situs web yang membahas mengenai ekspresi idiomatik bahasa Indonesia yang berjumlah 104 halaman web dengan total jumlah klaim 4358 klaim. Data ini akan digunakan pada tahap validasi *Truth Discovery*. Contoh data ditunjukkan pada Tabel 4.

Tabel 4. Contoh Data Situs Web

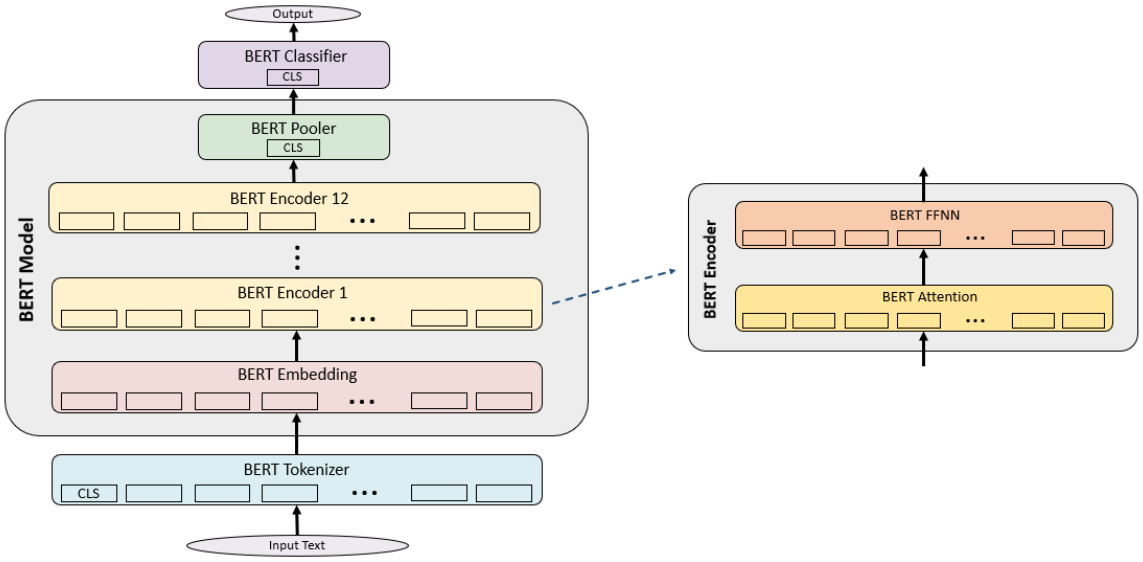
indeks	source_web	source_url	idiom_claims
1	salamadian.com	https://salamadian.com/contoh-ungkapan-bahasa-indonesia-dan-artinya-idiom/	Adu Mulut
2	salamadian.com	https://salamadian.com/contoh-ungkapan-bahasa-indonesia-dan-artinya-idiom/	Akal Bulus
3	salamadian.com	https://salamadian.com/contoh-ungkapan-bahasa-indonesia-dan-artinya-idiom/	Angkat Bicara
4	salamadian.com	https://salamadian.com/contoh-ungkapan-bahasa-indonesia-dan-artinya-idiom/	Angkat Kaki

2.3 *Text Preprocessing*

Pada penelitian ini, data kumpulan kalimat bahasa Indonesia akan dilakukan *text preprocessing* terlebih dahulu agar data tersebut siap dan dapat diolah untuk tahap selanjutnya. *Text preprocessing* digunakan untuk menyajikan data berupa teks dalam format yang sesuai. Adapun langkah-langkah yang dilakukan untuk *text preprocessing* pada penelitian ini adalah *punctuation removal* untuk menghilangkan simbol-simbol yang tidak diperlukan pada kalimat [7], *tokenization untuk memecah kalimat menjadi token yang dalam hal ini berupa kata*, dan *case conversion* untuk mengkonversi bentuk huruf dalam teks menjadi seragam yaitu menjadi huruf kecil [8].

2.4 *Bidirectional Encoder Representations from Transformers* (BERT)

Pada tahap ini dilakukan klasifikasi terhadap kalimat inputan ke dalam dua kategori yaitu 'kalimat_biasa' jika kalimat tidak mengandung ekspresi idiomatik dan 'kalimat_idiom' jika kalimat mengandung ekspresi idiomatik. Tahap ini membutuhkan model klasifikasi yang dibangun menggunakan BERT. *Bidirectional Encoder Representations from Transformers* atau disingkat BERT adalah model representasi bahasa terlatih yang dikembangkan oleh Devlin et al. (2019) [9]. BERT merupakan metode *state-of-the-art* dalam pembangunan *language model* dengan pendekatan *deep learning*. Arsitektur BERT dapat dilihat pada gambar berikut [10].



Gambar 2. Arsitektur BERT

2.5 POS *Tagging*

Pada penelitian ini, POS *Tagging* digunakan untuk memberikan label pada setiap kata dalam kalimat dengan kelas kata yang sesuai untuk kata tersebut, seperti kata benda, kata kerja, kata sifat, dan lain-lain. Pada tahap POS *Tagging* ini membutuhkan model yang dibangun menggunakan algoritma *Hidden Markov Model* (HMM). Adapun persamaan HMM ditunjukan pada persamaan (1) [11]:

$$t_1^n = \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{emisi} \overbrace{P(t_i|t_{i-1})}^{transisi} \quad (1)$$

Dengan $P(t_i|t_{i-1})$ merupakan probabilitas transisi yang mewakili probabilitas sebuah tag jika diketahui tag sebelumnya, yang dapat dihitung dengan persamaan (2):

$$P(t_i|t_{i-1}) = \frac{Count(t_{i-1}, t_i)}{Count(t_{i-1})} \quad (2)$$

Dan $P(w_i|t_i)$ merupakan probabilitas emisi yaitu probabilitas sebuah kata yang dilabeli *tag* tertentu, yang dihitung dengan persamaan (3).

$$P(w_i|t_i) = \frac{Count(t_i, w_i)}{Count(t_i)} \quad (3)$$

Keterangan:

t_1^n = kelas kata yang dicari

p-ISSN: 2301-5373
e-ISSN: 2654-5101

w_i = kata yang dicari kelas katanya

t_i = kelas kata dari w_i yang ada di *corpus*

t_{i-1} = kelas kata sebelum kelas kata dari w_i yang ada di *corpus*

Part-of-Speech Tagging pada penelitian ini dilakukan untuk memudahkan tahap selanjutnya yaitu *chunking* dimana kata-kata akan dikelompokan ke dalam bentuk frasa yang ditentukan berdasarkan label kelas kata dari kata-kata tersebut.

2.6 *Chunking*

Pada penelitian ini, metode *chunking* digunakan untuk menemukan frasa nomina, frasa verba ataupun frasa adjektiva yang akan dikategorikan sebagai frasa kandidat dimana frasa tersebut memiliki konstruksi sintaksis pembentuk ekspresi idiomatik. Untuk menemukan frasa kandidat tersebut pada suatu kalimat, penulis akan mendefinisikan tata bahasa *chunk* (*chunk grammar*) menggunakan *tag* dari *part-of-speech tagging*, yang terdiri dari aturan *Regular Expressions* yang menunjukkan bagaimana kalimat harus dipotong.

2.7 *Distributional Semantic Based Approach*

Pada tahap ini akan dilakukan identifikasi terhadap frasa-frasa kandidat yang dihasilkan dari proses *chunking* apakah frasa tersebut merupakan ekspresi idiomatik atau tidak menggunakan pendekatan berbasis semantik distribusi (*Distributional Semantic Based Approach*) [4] [3] dengan menjumlahkan kemiripan semantik antara kandidat dan kumpulan contoh ekspresi idiomatik dengan kemiripan semantik antara kata penyusun frasa kandidat yang dapat dilihat pada persamaan (4):

$$sim = \sum_{i=1}^{n} sim(phrase, idiom\ example_i) + sim(w_1, w_2) \qquad (4)$$

Keterangan:
sim = *similarity*
$phrase$ = frasa kandidat idiom
$idiom\ example$ = contoh idiom
w_1, w_2 = kata-kata penyusun frasa kandidat idiom

Pada tahap ini, perhitungan nilai *similarity* dilakukan dengan dengan menggunakan *cosine similarity* yang menerima inputan berupa representasi vektor dari kata-kata dan juga frasa yang dihasilkan dari BERT *Embedding*. Setelah mendapatkan hasil nilai *similarity* antar frasa kandidat dengan contoh idiom dan nilai *similarity* antar kata penyusun frasa kandidat, selanjutnya kedua nilai *similarity* tersebut dijumlahkan berdasarkan persamaan (4), kemudian akan dipilih dari *frasa-frasa* kandidat tersebut yang diidentifikasi sebagai frasa idiom berdasarkan frasa yang memiliki nilai *similarity* lebih besar dari batas yaitu 0,5.

2.8 *Truth Discovery*

Truth Discovery atau bisa disebut pengecekan fakta bertugas untuk menemukan pernyataan yang benar diantara banyaknya pernyataan yang diklaim yang diajukan banyak sumber untuk objek yang sama [12]. Pada penelitian ini, proses validasi otomatis frasa idiom menggunakan kombinasi algoritma *Sums* dan *Average-Log* yang persamaannya ditunjukan oleh (5), (6) untuk algoritma Sums dan (7) untuk algoritma *Average-Log* [13].

$$T^i(s) = \sum_{c \in C_s} B^{i-1}(c) \qquad (5)$$

$$B^i(c) = \sum_{s \in S_c} T^i(s) \qquad (6)$$

$$T^i(s) = \log|C_s| \cdot \frac{\sum_{c \in C_s} B^{i-1}(c)}{|C_s|} \qquad (7)$$

Keterangan:

S = kumpulan sumber

S_c = kumpulan sumber yang memberikan klaim c

C = kumpulan klaim

C_s = kumpulan klaim disediakan oleh sumber s

$T^i(s)$ = *trustworthiness score* dari sumber s

$B^i(c)$ = *belief score* dari klaim c

$B^{i-1}(c)$ = *belief score* sebelumnya dari klaim c

Algoritma *Sums* dan *Avarage-Log* akan menghasilkan *trustworthiness score* dari sumber *website* yang dihitung menggunakan kombinasi persamaan (5) dan (7) dan *belief score* dari setiap klaim yang diberikan oleh sumber yang dihitung menggunakan persamaan (7). Kemudian untuk memvalidasi suatu frasa merupakan frasa idiom akan menggunakan *belief score*, dimana suatu frasa dianggap benar atau valid sebagai frasa idiom jika mempunyai *belief score* di atas nilai parameter *threshold* (t) yaitu persentil 15 dari data *belief score* yang persamaanya ditunjukkan oleh (5).

$$t = percentile(belief\ score\ , 15) \quad (8)$$

Langkah-langkah melakukan validasi otomatis menggunakan *Truth Discovery* antara lain:

a. Menginput data sumber web dan data frasa bahasa Indonesia sebagai dataset, *initial value* yang ditetapkan pada penelitian ini yaitu 0.5 dan *iteration value* adalah 3.
b. Membuat tabel *One Hot Encoding* yang berisikan *claim value* c pada sumber s, dimana *claim value* c akan bernilai 1 jika klaim c tersedia pada sumber s dan bernilai 0 untuk sebaliknya.
c. Menghitung nilai c*laim source score* C_S yaitu banyaknya jumlah klaim yang diberikan oleh setiap sumber s.
d. Melakukan inisialisasi *claim value* = *claim value* x *initial value.*
e. Menghitung *trustworthiness score* sumber s dengan menggunakan persamaan (5) dan (7).
f. Menghitung *belief score* klaim c dengan menggunakan persamaan (6)
g. Mengulangi langkah e dan f sesuai *iteration value*.
h. Menghitung nilai parameter *threshold* dengan menggunakan persamaan (8)
i. Melakukan proses validasi dimana suatu frasa dianggap benar atau valid sebagai frasa idiom jika mempunyai *belief score* di atas nilai parameter *threshold*.

2.9 Evaluasi Sistem

Pada penelitian ini akan dilakukan pengukuran performa menggunakan *confusion matrix*. Parameter yang digunakan adalah *accuracy* (9), *precision* (10), *recall* (11) dan *f1-score* (12).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FP} \quad (11)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

Keterangan:

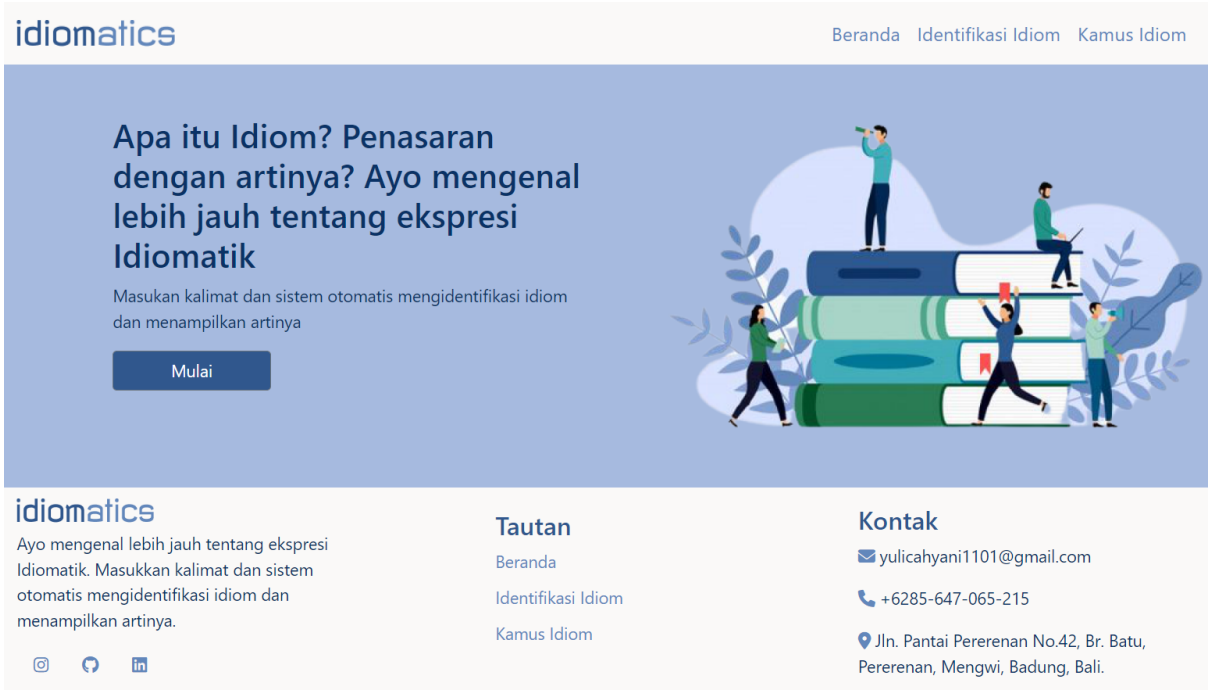
a. *True Positive* (TP), yaitu total hasil dari prediksi kelas positif dan sesuai dengan kelas aslinya yang positif.
b. *True Negative* (TN), yaitu total hasil dari prediksi kelas negatif dan sesuai dengan kelas aslinya yang negatif.
c. *False Positive* (FP), yaitu total hasil dari prediksi kelas positif namun tidak sesuai dengan kelas aslinya yang negatif.
d. *False Negative* (FN), yaitu total hasil dari prediksi kelas negatif namun tidak sesuai dengan kelas aslinya yang positif.

3. Hasil dan Pembahasan

3.1. Tampilan Antarmuka Sistem Identifikasi Ekspresi Idiomatik

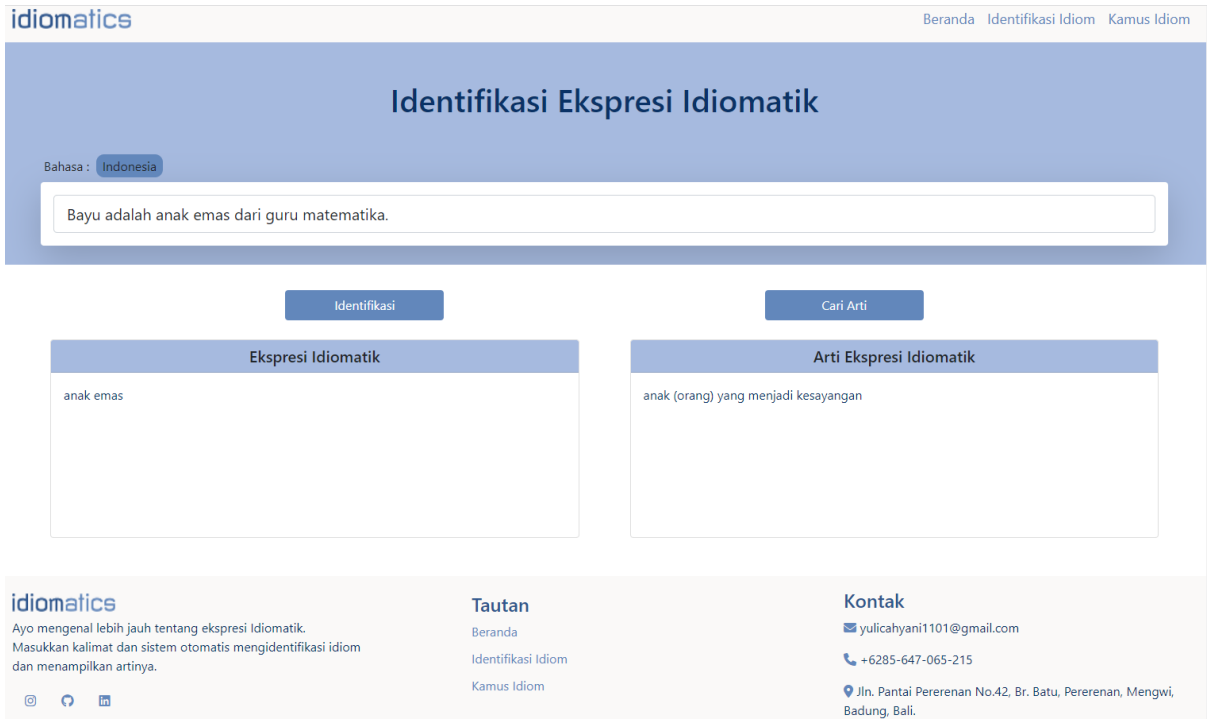
Tampilan antarmuka sistem terdiri dari tiga halaman, yaitu halaman beranda, halaman identifikasi idiom, dan halaman kamus idiom. Pada saat pertama kali membuka sistem, akan ditampilkan halaman beranda yang dapat dilihat pada Gambar 3. Pada halaman beranda menampilkan nama sistem serta deskripsi singkat dari sistem.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 3. Halaman Beranda

Halaman identifikasi yang dapat dilihat pada Gambar 4. merupakan halaman untuk melakukan identifikasi ekspresi idiomatik. Pada halaman ini terdapat *text box* untuk menerima inputan dari user yang berupa suatu kalimat berpola dasar berbahasa Indonesia. Kemudian terdapat tombol identifikasi untuk melakukan proses identifikasi ekspresi idiomatik pada kalimat inputan dan juga terdapat tombol cari arti untuk menampilkan arti dari ekspresi idiomatik yang telah diidentifikasi sebelumnya.



Gambar 4. Halaman Identifikasi Idiom

Selanjutnya halaman kamus idiom dapat dilihat pada Gambar 5. Pada halaman kamus idiom terdapat *text box* untuk menerima inputan dari user yang berupa huruf, kata, atau frasa berbahasa Indonesia. Kemudian terdapat tombol cari untuk melakukan proses pencarian berdasarkan inputan dan menampilkan hasil berupa ekspresi idiomatik beserta arti dan contoh penggunaanya dalam kalimat.

Gambar 5. Halaman Kamus Idiom

3.2. Evaluasi Identifikasi Ekspresi Idiomatik

Pada penelitian ini, untuk mengetahui performa identifikasi ekspresi idiomatik menggunakan *Distributional Semantic Based Approach* dan *Truth Discovery* dilakukan pengujian atau evaluasi untuk mendapatkan nilai dari akurasi, presisi, *recall*, dan *f1- score*. Pada pengujian ini, seluruh data kalimat Bahasa Indonesia digunakan sebagai data uji, hal ini terjadi karena dalam identifikasi yang tidak memerlukan data latih. Pengujian akan dilakukan sebanyak 10 kali dengan jumlah data uji yang berbeda-beda dalam setiap pengujian. Selanjutnya hasil evaluasi didapatkan melalui rata-rata dari 10 kali pengujian tersebut. Hasil evaluasi identifikasi ekspresi idiomatik menggunakan *Distributional Semantic Based Approach* dan *Truth Discovery* dapat dilihat pada Tabel 5.

Tabel 5. Hasil Evaluasi Identifikasi Ekspresi Idiomatik

Pengujian	Jumlah Data	Akurasi	Presisi	*Recall*	*F1-Score*
1	200	0.83	1.0	0.67	0.80
2	400	0.82	1.0	0.62	0.76
3	600	0.82	1.0	0.66	0.79
4	800	0.82	1.0	0.63	0.77
5	1000	0.82	1.0	0.65	0.79
6	1200	0.82	1.0	0.64	0.78
7	1400	0.83	1.0	0.66	0.79
8	1600	0.82	1.0	0.64	0.78
9	1800	0.81	1.0	0.64	0.78
10	2000	0.82	1.0	0.64	0.78
Rata-rata		**0.82**	**1.0**	**0.64**	**0.78**

Pada Tabel 5. di atas dapat kita lihat nilai akurasi, presisi, *recall* dan *f1-score* yang diperoleh dari pengujian dengan jumlah data 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800 dan 2000. Dari pengujian yang telah dilakukan, dapat dikatakan model identifikasi ekspresi idiomatik menggunakan *Distributional Semantic Based Approach* dan *Truth Discovery* memiliki performa cukup baik dengan rata-rata nilai akurasi, presisi, *recall* dan *f1-score* yang diperoleh yaitu 0,82; 1,0; 0,64 dan 0,78. Rata-rata akurasi, presisi dan *f1-score* memiliki nilai yang cukup baik sedangkan nilai *recall* yang dihasilkan

cukup rendah, yang berarti jumlah ekspresi idiomatik yang tidak berhasil diidentifikasi cukup banyak. Hal tersebut dapat disebabkan oleh beberapa faktor sebagai berikut:

a. Pada tahap klasifikasi, BERT tidak berhasil mengklasifikasikan kalimat ke dalam kategori kalimat_idiom.
b. Prediksi kelas kata yang dihasilkan pada tahap *Part-of-Speech Tagging* kurang optimal sehingga menyebabkan proses *Chunking* juga memberikan hasil yang kurang optimal dalam menemukan frasa kandidat yang memiliki konstruksi sintaksis pembentuk ekspresi idiomatik.
c. Keterbatasan data sumber web, dimana klaim berupa ekspresi idiomatik yang disediakan oleh sumber web cenderung sedikit sehingga pada proses identifikasi terdapat beberapa frasa yang sebenarnya merupakan idiom tapi tidak dinyatakan sebagai idiom karena tidak ada klaim dari sumber web atas frasa tersebut.

4. Kesimpulan

Berdasarkan pada penelitian yang telah dilakukan serta hasil yang diperoleh dari 10 kali pengujian dengan jumlah data yang berbeda menunjukkan bahwa identifikasi ekspresi idiomatik menggunakan *Distributional Semantic Based Approach* dan *Truth Discovery* memiliki performa yang cukup baik dalam mengidentifikasi ekspresi idiomatik pada kalimat berbahasa Indonesia dengan rata-rata akurasi sebesar 0,82; presisi sebesar 1,0; *recall* sebesar 0,64 dan *f1-score* sebesar 0,78. Dari penelitian yang telah dilakukan serta hasil yang diperoleh, saran-saran yang dapat disampaikan untuk dapat dipertimbangkan dalam pengembangan dari penelitian selanjutnya yaitu Algoritma HMM untuk *Part-of-Speech Tagging* dapat diganti dengan algoritma lainnya seperti *Brill Tagger* yang menggunakan aturan leksikal dan kontekstual berdasarkan *Transformation Based Learning* untuk mendapatkan hasil prediksi kelas kata yang lebih optimal. Kemudian memperbanyak data sumber web yang digunakan dalam membangun model *Truth Discovery* sehingga dalam identifikasi ekspresi idiomatik memperoleh akurasi, presisi, *recall* dan *f1-score* yang lebih baik.

Daftar Pustaka

[1] V. V. Virdaus, "Ekspresi Idiomatik dalam Lirik Nine Track Mind Charlie Puth Album 2016", *Media of Teaching Oriented and Children*, vol.4, no.1, 2020.

[2] A. Barrera, R. Verma and R. Vincent, "SemQuest: University of Houston's Semantics-based Question Answering System" in *Text Analysis Conference (TAC)*, Houston, Nist.Gov, 2011.

[3] O. Zayed, J. P. McCrae, and P. Buitelaar, "Phrase-level Metaphor Identification using Distributed Representations of Word Meaning," in *Proceedings of the Workshop on Figurative Language Processing*, New Orleans, Louisiana, 2018, pp. 81–90.

[4] E. Shutova, D. Kiela, J. Maillard, "Black Holes and White Rabbits: Metaphor Identification with Visual Features" in *Proceedings of NAACL-HLT*, San Diego, California, 2016, pp.160–170.

[5] A. Chaer, *Pengantar Semantik Bahasa Indonesia*, Jakarta: Rineka Cipta, 2013.

[6] A. Anggito, and J. Setiawan, *Metodelogi Penelitian Kualitatif*, Jawa Barat: CV Jejak, 2018.

[7] D. Sarkar, *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second ed.*, Bangalore, Karnataka, India: Appress Media, 2019.

[8] A. E. Karyawati, P. A. Utomo, and I. G. A. Wibawa, "Comparison of SVM and LIWC for Sentiment Analysis of SARA," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 16, no. 1, p. 45, 2022, doi: 10.22146/ijccs.69617.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding" in *NAACL HLT (Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*, Minneapolis, Minnesota, 2019, vol.1, pp. 4171–4186.

[10] S. Kachuee and M. Sharifkhani, "TiltedBERT: Resource Adjustable Version of BERT," 2022, [Online]. Available: http://arxiv.org/abs/2201.03327.

[11] D. Jurafsky, and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Third ed.*, Palo Alto: Stanford University, 2017.

[12] N. A. Sanjaya, T. Abdessalem, M. L. Ba, and S. Bressan, “Harnessing Truth Discovery Algorithms on The Topic Labelling Problem”, in *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*, 2018, pp. 8-14.

[13] J. Pasternack and D. Roth, “Knowing What to Believe (when you already know something)”, in *Coling (International Conference on Computational Linguistics)*, 2010, vol. 2, pp. 877–885.